<Grant Agreement Number>

EuropeanaLocal

D4.5 Directory of vocabularies converted through SKOS

Deliverable number	*D-4.5*
Dissemination level	*Open*
Delivery date	*M29*
Status	*Final*
Author(s)	*(Stein) Runar Bergheim, Avinet* *Olav Tuften, Avinet* *Rastislav Rehak, EEA*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Table of Contents

1 Introduction

This document serves as partial evidence towards the the completion of Work Package 4 tasks in the Description of Work for the EuropeanaLocal project. A brief introduction to the underlying work of this list is given below.

Throughout the EuropeanaLocal project, it became evident that few of the heritage institutions contributing content to Europeana systematically enforce the use of controlled vocabulariesin their applications. Many institutions, and indeed countries, have defined vocabularies in printed form, but the practical implementation of these are mostly as free-text fields in a registration form where it is at the discretion of the person cataloging the content to select the keywords and ensure consistency throughout her or his work.

This practice, while sufficient in a single source environment, leads to challenges once the content is integrated and mixed with other sources where purely textual keywords, without any definition, may have different meaning dependant on the context.

Chapter 3 below is the “directory” itself and provides an overview of widely used “international” controlled vocabularies, thesauri, available as RDF/SKOS which are suitable to be mapped to from local, single institutions

4 below provides a brief overview of the issues which should be covered throughout.

2 Glossary

Term	Meaning
XML	eXtensible Mark-up Language
URI	Universal Resource Identifier
HTTP	Hyper-Text Transfer Protocol
SKOS	Simple Knowledge Organization System
Thesaurus	Plural: thesauri,

3 Directory of vocabularies converted through SKOS

3.1 UNESCO

Concept scheme: http://iaaa.unizar.es/thesaurus/UNESCO

3.1.1 Description

The UNESCO Thesaurus is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for the following areas of knowledge: education, science, culture, social and human sciences, information and communication, and politics, law and economics. It also includes the names of countries and groupings of countries: political, economic, geographic, ethnic and religious, and linguistic groupings.

3.1.2 Usage

The UNESCO Thesaurus allows subject terms to be expressed consistently, with increasing specificity, and in relation to other subjects. It can be used to facilitate subject indexing in libraries, archives and similar institutions.

3.1.3 Metadata

Number of rdf:Description elements	4425
Number of labels	13257
Number of broader term relations	4833
Number of narrower term relations	4833
Number of related term relations	12015

3.1.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- FORCED LABOUR
- SOCIO-ECONOMIC ANALYSIS
- FINE ARTS
- MARRIED WOMEN
- FALKLAND ISLANDS
- SKIN DISEASES
- HYDROELECTRIC POWER
- SOIL MAPS

- PARASITOLOGY
- ISBD
- CATALOGUING
- INFECTIOUS DISEASES
- SCIENTIFIC COOPERATION
- TECHNOLOGY TRANSFER
- URANIUM
- DYSLEXIA
- EDUCATIONAL EFFICIENCY
- DOMESTIC WORKERS
- TRADE MARKS
- RELIGIOUS BUILDINGS
- REMOTE SENSING
- THEATRICAL PRODUCTION
- PERIODICAL PRESS
- FURTHER TRAINING
- SCHOOL COMMUNITY RELATIONSHIP
- REPETITION RATE
- MARINE ANIMALS
- TYPOLOGY
- BRAIN DRAIN
- ISBN
- DEAD SEA
- RELIGION
- FAUNA
- PILOT PROJECTS
- NEWSLETTERS
- DEVELOPMENT PLANS
- CHEMICAL RESEARCH
- GENERAL TECHNICAL EDUCATION
- INDUSTRY
- DEMOCRATIZATION OF CULTURE
- VALUE SYSTEMS
- NEW CALEDONIA
- PEDIATRICS
- COMPOSITE MATERIALS
- DOCUMENTARY INFORMATION PROCESSING
- PROJECT DESIGN
- AGRICULTURAL RESEARCH
- CULTURAL BEHAVIOUR
- STUDENT MOVEMENTS
- MOSLEMS

3.2 Library of Congress Subject Headings

Concept scheme: http://id.loc.gov/authorities#conceptscheme

3.2.1 Description

Library of Congress Subject Headings (LCSH) has been actively maintained since 1898 to catalog materials held at the Library of Congress. By virtue of cooperative cataloging other libraries around the United States also use LCSH to provide subject access to their collections. In addition LCSH is used internationally, often in translation. LCSH in this service includes all Library of Congress Subject Headings, free-floating subdivisions (topical and form), Genre/Form headings, Children's (AC) headings, and validation strings* for which authority records have been created. The content includes a few name headings (personal and corporate), such as William Shakespeare, Jesus Christ, and Harvard University, and geographic headings that are added to LCSH as they are needed to establish subdivisions, provide a pattern for subdivision practice, or provide reference structure for other terms. This content is expanded beyond the print issue of LCSH (the "red books") with inclusion of validation strings.

3.2.2 Usage

Used in modified form by several Europeana contributors.

3.2.3 Metadata

Number of rdf:Description elements	406647
Number of labels	742641
Number of broader term relations	260071
Number of narrower term relations	260071
Number of related term relations	22046

3.2.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- Mayoruna language
- Nambicuara language
- Hixkaryana language
- Wayana language
- Rikbaktsa language
- Sharanahua language
- Siriano language
- Trio language
- Trumai language
- Tucuna language
- Waiwai language
- Urubu language
- Tanimuca-Retuama language

- Yanomamo language
- Yecuana language
- Canamari language (Tucanoan)
- Kraho language
- Piratapuyo language
- Mura language
- Münkü dialect
- Pirahá dialect
- Cashinawa language
- Cayapo language
- Brazil--Politics and government
- Coronelismo
- Brazil--Politics and government--1763-1822
- Brazil--Politics and government--1822-1889
- Brazil--Politics and government--1822-
- Brazil--Politics and government--1889-1930
- Brazil--Politics and government--1889-
- Brazil--Politics and government--20th century
- Brazil--Politics and government--1930-1945
- Idaho Panhandle (Idaho)
- Brazil--Politics and government--1954-1964
- Brazil--Politics and government--1964-1985
- Brazil--Politics and government--1985-2002
- Palácio Itamaraty (Brasília, Brazil)
- Palácio das Indústrias (São Paulo, Brazil)
- Palácio da Liberdade (Belo Horizonte, Brazil)
- Brazil--Social conditions
- Brazil--Social conditions--19th century
- Brazil--Social conditions--1945-1964
- Brazil--Social conditions--1964-1985
- Brazil--Social life and customs
- Brazil--Social life and customs--19th century
- Brazil--Social life and customs--20th century
- Brazil, Central West
- Brazil, North
- Brazil, Northeast
- Brazil, Northeast--History

3.3 AGROVOC

Concept scheme: http://iaaa.unizar.es/thesaurus/AGROVOC

3.3.1 Description

Specialized thesaurus for the classification of geographic information resources (with special focus on agriculture resources) has been created by the Food and Agriculture Organization of the United Nations (FAO). It is distributed free of charge and is available in 7 different languages: Arabic, Chinese, English, French, Spanish, Czech and Portuguese.

3.3.2 Usage

Useful for classification of agricultural resources

3.3.3 Metadata

Number of rdf:Description elements	12655
Number of labels	57436
Number of broader term relations	11468
Number of narrower term relations	11330
Number of related term relations	19718

3.3.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- Chlorsulfuron
- Chronic toxicity
- Clenbuterol
- Carbon cycle
- Carbosulfan
- Chlorimuron
- Chlorination
- Coleus
- Coleus amboinicus
- Cold
- Cold stores
- Cold zones
- Colchicum
- Collectivization
- Collective farming
- Nature conservation and land resources
- Colistium
- Collagen
- Coliiformes
- Renewable energy resources
- Energy resources and management
- Coleus rotundifolius
- Non-renewable energy resources
- Colic
- Fluroxypyr
- Fluvalinate
- Land degradation

- Fenoxaprop
- Fibrinogen
- Environmental degradation
- Colocasia esculenta
- Cololabis
- Drainage
- Water resources and management
- Collenchyma
- Colletotrichum
- Colloidal properties
- Colloids
- Colocasia
- Entomogenous bacteria
- Eicosanoids
- Diesel oil
- Cysts
- Diameter increment
- Compensatory growth
- Commercial banks
- Commercial farming
- Combretaceae
- Columbiformes
- Combine harvesters

3.4 GEMET General Multilingual Environmental Thesaurus

Concept scheme: http://www.eionet.europa.eu/gemet

3.4.1 Description

GEMET, the GEneral Multilingual Environmental Thesaurus, has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen. The work has been carried out through a contract between the EEA and the ETC/CDS which is led by the Ministry of the Environment of Lower Saxony, includes members of Germany, Austria, Italy, Sweden and benefits of the collaboration of other member countries of the European Union (EU), as well as of UNEP Infoterra. The basic idea for the development of GEMET was to use the best of the presently available excellent multilingual thesauri, in order to save time, energy and funds. GEMET was conceived as a “general” thesaurus, aimed to define a common general language, a core of general terminology for the environment. Specific thesauri and descriptor systems (e.g. on Nature Conservation, on Wastes, on Energy, etc.) have been excluded from the first step of development of the thesaurus and have been taken into account only for their structure and upper level terminology.

3.4.2 Usage

While fundamentally an environmental thesaurus, the use is currently being evaluated within Europeana due to the massive language support - a feature which is unrivalled in any other current thesaurus applicable to the European Union area.

3.4.3 Metadata

Number of rdf:Description elements	5208
Number of labels	5208
Number of broader term relations	5191
Number of narrower term relations	5191
Number of related term relations	2086

3.4.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- abandoned industrial site
- abandoned vehicle
- abiotic factor
- absorption (exposure)
- acceptable risk level
- agreement (administrative)
- access road
- access to culture
- access to the sea
- accident
- accidental release of organisms
- accident source
- accumulation in body tissues
- accumulator
- acid deposition
- acidification
- acidity
- acidity degree
- acid rain
- acid
- acoustic filter
- acoustic insulation
- acoustic level
- acoustic property
- acoustics
- actinide
- actinium
- action group
- activated carbon
- activated sludge
- active participation

- intervention on land
- act
- adaptable species
- adaptation period
- chemical addition
- additional packaging
- addition polymer
- additive
- adhesive
- acceptable daily intake
- administration
- administrative body
- administrative competence
- administrative fiat
- administrative jurisdiction
- administrative law
- administrative procedure
- administrative sanction
- adsorption

3.5 ISO639 - Language Codes

Concept scheme: http://iaaa.unizar.es/thesaurus/ISO639

3.5.1 Description

Controlled vocabulary that contains the spoken languages in the world. It contais the ISO639-1, ISO639-2 e ISO639-3 standards. This list contains the name of the language and the 2 and 3 letter codes. It uses the 2006 version.

3.5.2 Usage

Widely used throughout Europeana contributors for specifying the language of metadata and resources

3.5.3 Metadata

Number of rdf:Description elements	7600
Number of labels	45701
Number of broader term relations	0
Number of narrower term relations	0
Number of related term relations	0

3.5.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- Quechua, Napo Lowland
- Quechua, North Junín
- Quechua, Pacaraos
- Quichua, Loja Highland
- Quechua, Margos-Yarowilca-Lauricocha
- Quechua, Cajatambo North Lima
- Quechua, Huaylla Wanca
- Queyu
- Quechua, San Martín
- Quechua, Ambo-Pasco
- Quechua, Ayacucho
- Quechua, Cusco
- Quechua, Huamalíes-Dos de Mayo Huánuco
- Quichua, Imbabura Highland
- Quechua, Cajamarca
- Quechua, Eastern Apurímac
- Quechua, Southern Pastaza
- Quinault
- Sipacapense
- Quechua, North Bolivian
- Quechua, Chachapoyas
- Quiché, Joyabaj
- Quileute
- Quechua, Yauyos
- Quichua, Tena Lowland
- Sacapulteco
- Quiché, Eastern
- Quiché, West Central
- Quichua, Santiago del Estero
- Quechua, Yanahuanca Pasco
- Quinqui
- Quichua, Chimborazo Highland
- Quechua, South Bolivian
- Quechua
- Quechua, Lambayeque
- Quiché, Central
- Quichua, Calderón Highland
- Quapaw
- Quechua, Huallaga Huánuco
- Gapapaiwa
- Pawaia
- Karen, Pwo Northern
- Molbog
- Paiwan
- Karen, Pwo Western

- Powari
- Mixe, Quetzaltepec
- Phuie
- Punan Merah
- Puinave

3.6 Thesaurus for the Social Sciences (TheSoz)

Concept scheme: http://lod.gesis.org/thesoz/

3.6.1 Description

The Thesaurus for the Social Sciences (Thesaurus Sozialwissenschaften) contains about 11,600 entries, of which more than 7,750 are descriptors (authorised keywords) and about 3,850 non-descriptors. Topics in all of the social science disciplines are included. This SKOS version of the thesaurus uses also SKOS-XL and additionally defined extensions.

3.6.2 Usage

3.6.3 Metadata

Number of rdf:Description elements	158197
Number of labels	410
Number of broader term relations	17748
Number of narrower term relations	17749
Number of related term relations	3364

3.6.4 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- Thesaurus for the Social Sciences (TheSoz)
- Fundamentals of the Social Sciences
- Philosophy of Science, Methodology, Methods
- Philosophy of Science, Methodology
- Types of Research, Research Design
- Data Collecting and Analysis Techniques, Statistics
- Planning and Decision-Making Techniques (Implementation)
- Theories and Theoretical Approaches
- Labor Market and Occupational Research
- Science of Communication
- Pedagogics
- Philosophy

- Political Science
- Psychology
- Jurisprudence
- Sociology/ Social Psychology
- Economics
- Other Sciences
- Scientific Disciplines and Subsections
- Labor Market and Occupational Research
- Science of Communication
- Pedagogics
- Philosophy
- Political Science
- Psychology
- Jurisprudence
- Sociology/ Social Psychology
- Economics
- Other Disciplines and Subsections
- Society
- Social System, Social Stratification
- Social System
- Social Stratification
- Social Class
- Power, Domination
- Social Mobility
- Social Inequality
- Social Environment
- Organization and Institution
- Structure of Society, Social Movements, Ideologies
- Structure of Society
- Social Movements
- Ideologies
- Social Change
- Social Change
- Integration and Segregation
- Innovation and Diffusion
- Social Data and Social Indicators
- Fundamentals and Manifestations of Social Behavior
- Individual, Personality

Concept scheme:

3.6.5 Description

3.6.6 Usage

3.6.7 Metadata

Number of rdf:Description elements 36154

Number of labels	12439
Number of broader term relations	7364
Number of narrower term relations	7372
Number of related term relations	13802

3.6.8 Sample content

The following sample consists of the first 50 labels present in the thesaurus XML-file and are meant as a sample of the type of content and level of detail present in the data.

- 4-H clubs
- A la poupée prints
- A trois crayons drawings
- Abacus
- Abandoned buildings
- Abandoned children
- Abandoned farms
- Abandoned mines
- Abbeys
- Abdication
- Ablution fountains
- Abolition movement
- Abolitionists
- Abortions
- Absent mindedness
- Absenteeism (Labor)
- Abstract drawings
- Abstract paintings
- Abstract photographs
- Abstract prints
- Abstract sculpture
- Abstract works
- Abused children
- Abused women
- Abutments
- Abyss
- Acanthi
- Accidents
- Accordions
- Acetate negatives
- Acolytes
- Acorn decorations
- Acoustical engineering
- Acrobatics
- Acrobats

- Acrylic paintings
- Action & adventure dramas
- Actions & defenses
- Activists
- Activities
- Actors
- Actresses
- Acupuncture
- Acupuncture anesthesia
- Adaptive reuse
- Adits
- Administrative agencies
- Admirals
- Adobe buildings
- Adobe houses

4 Summary and Conclusions

The systematic use of controlled vocabularies is a way to secure horizontal integration of Europeana content from different content providers.

By applying controlled vocabularies in a consistent manner – and mapping concepts from local vocabularies to corresponding concepts in widely used “international“ vocabularies such as those listed in this deliverable, a comprehensive graph will be built. This graph will connect content from different content providers and make it possible to relate e.g. concepts from a local collection in Norway to those of a local collection in Bulgaria.

Currently, few cultural heritage institutions are enforcing the use of controlled vocabularies, and for these to comply with the regulations a lot of semi-automatic or manual work is required.

Progress is being made in the domain of auto-extraction whereby metadata or corpus text is processed for occurences of concepts from controlled vocabularies which are then automatically added to the metadata. This would offer a more cost-effective, but less precise, way of “upgrading” existing collections to support controlled vocabularies.

SKOS is one of the prime technologies for expressing controlled vocabularies in XML and is suitable both for input to cataloguing applications and for real-time auto-extraction tools, as described above.

Where EuropeanaLocal ends, there is excellent potential for partners to work further on refining their metadata and content – topics which have been covered in great detail throughout technical workshops and joint partner meetings throughout the project.